Adjacency and Proximity Searching in the Science Citation Index and Google

Dr. Ronald N. Kostoff
Office of Naval Research, 875 N. Randolph St, .Arlington, VA 22217

Mr. John T. Rigsby
Naval Surface Warfare Center Dahlgren Division, 17320 Dahlgren Road B10, Dahlgren, VA 22448

Mr. Ryan B. Barth
DDL-OMNI Engineering, LLC, 8260 Greensboro Drive, McLean, VA 22102

Correspondence to: Dr. Ronald N. Kostoff, Office of Naval Research, 875 N. Randolph St., Arlington, VA 22217; kostofr@onr.navy.mil

Abstract

We have developed simple algorithms that allow adjacency and proximity searching in Google and the Science Citation Index (SCI).

The SCI algorithm exploits the fact that SCI stopwords in a search phrase function as a placeholder. Such a phrase serves effectively as a fixed adjacency condition determined by the number n of adjacent stopwords (i.e., retrieve all records where word A and word B are separated by n words in at least one location). The algorithm integrates over search phrases with different numbers of adjacent stopwords to provide a flexible adjacency or proximity capability (i.e., retrieve all records where word A and word B are separated by n or less words in at least one location, where n is the maximum separation desired between A and B in at least one location).

The Google algorithm exploits the fact that asterisks (in Google) separating words in a phrase function like word wildcards. The difference between two such phrases (the first phrase containing one less asterisk than the second phrase) serves effectively as a fixed adjacency or proximity condition, with the number of separating words equal to the number of asterisks in the first phrase. The algorithm integrates over these phrase differentials to provide a flexible adjacency or proximity capability (i.e., retrieve all records where word A and word B are separated by n or less words in at least one location, where n is the maximum separation desired between A and B in at least one location).

This document is based on the journal paper "RN Kostoff, JT Rigsby, and RB Barth. Adjacency and proximity searching in the Science Citation Index and Google. Journal of Information Science. 32. In Press.". Also, based on the report "RN Kostoff, JT Rigsby, and RB Barth. Adjacency and proximity searching in the Science Citation Index and Google. DTIC Technical Report Number ADA 442 888 (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA. 2005."

Keywords: Information Retrieval; Adjacency Searching; Proximity Searching; Constrained Co-Occurrence Searching; Science Citation Index; Google; Yahoo; Engineering Compendex; PubMed; OVID; Search Engine; Query

1. Background

Many types of research and development activities require full knowledge of prior and related art. Patent applications, literature reviews, technology watch [1], science and technology roadmaps [2], literature-based discovery [3], and the proper conduct of fundamental research require familiarity with the global science and technology (S&T) literature related to the problem of interest.

One method of accessing this global S&T literature is through searching large electronic databases. The operational information retrieval challenge is twofold: retrieve as many relevant records as possible (recall) from the appropriate databases, while at the same time insuring that a high fraction of the retrieved records is relevant (precision). In practice, some tradeoffs between recall and precision are usually required.

When performing keyword searching of these databases, using a combination of words that form a coherent phrase (e.g., 'information technology') typically provides maximum precision for that combination. However, recall is limited. Different authors may use variants of the same words/ phrases to describe the same phenomenon. Thus, while one author may use the expression 'nutrient uptake' to describe the ingestion of nutrients, another author may use 'uptake of nutrients' to describe the same phenomena. Limiting the search to records containing only the coherent phrase 'nutrient uptake' would ignore the many records containing the desired concept expressed in different terminology.

At the other extreme, requiring only co-occurrence of the word combination within a database record (e.g., 'information' AND 'technology'), or even co-occurrence in a free-text field within the record, provides maximum recall for that combination. In this case, precision could suffer, and many non-relevant records could be retrieved. The recall-precision tradeoff suggests that an interior optimization may be most desirable.

Constrained co-occurrence searching, sometimes called adjacency or proximity searching, operates between the adjacent word phrase and unconstrained co-occurrence extremes described above. Constrained co-occurrence searching allows records containing variant expressions of the desired target concept to be retrieved. Moreover, constrained co-occurrence searching allows the searcher to specify the desired spacing between the search words.

There are two major types of constrained co-occurrence searching: adjacency searching and proximity searching. Adjacency searching, as commonly practiced, is the identification of text records that contain words/ phrases

located within a specified range of each other in at least one location. Adjacency searching requires that the search words be in a specified order (e.g., 'information' must precede 'technology'), with either a fixed distance separating the words (e.g., 'information' precedes 'technology' by exactly three words) or within a specified range (e.g., 'information' precedes 'technology' by three words or less). McJunkin [4] examined the value of adjacency search on paper titles, and found that, when keywords were discipline specific, adjacency operators improved precision with little degradation in recall.

Proximity searching, as commonly used, does not specify any order relationship between the words, but only requires that the words occur within a specified range of each other (e.g., 'information' occurs within three words of 'technology'). Keen [5] performed a wide parametric sweep in examining proximity search techniques, and quantified the value of different proximity search approaches.

Thus, an adjacency search query term of the form A ADJ/4 B (applied to a given field(s)) is interpreted as 'retrieve all records in the database that contain at least one occurrence of word/ phrase A preceding, and within four words of word/ phrase B, in the field(s) specified'. A proximity search query term of the form A w/4 B (applied to a given field(s)) is interpreted as 'retrieve all records in the database that contain at least one occurrence of word/ phrase A within four words of word/ phrase B, in the field(s) specified'. In either case of constrained co-occurrence searching, setting the distance between the search words allows any desired balance between precision and recall (for a given combination of words) to be achieved. Having this user option for every major electronic database search engine of interest would be extremely desirable.

If one examines the major database search engines, including commercial S&T database search engines (e.g., Science Citation Index (SCI), Engineering Compendex (EC), PubMed, OVID), Federal agency award database search engines (e.g., NSF, NIH, DOE, EPA, as accessed in Federal R&D Project Summaries), Web search Engines (e.g., Google, Yahoo, Alta Vista), one finds major inconsistencies in all available search user options, including constrained co-occurrence searching. Some database search engines allow strict constrained co-occurrence searching as a user option (e.g., OVID, EC), while others do not (e.g., SCI, Google).

Perhaps the two search engines used most in our text mining work have been the SCI (for retrieving specific technical/ medical published journal articles in basic and applied research) and Google (for general purpose document retrieval). The absence of a strict constrained co-occurrence search user option for these databases has impacted our text mining efforts adversely.

This Communication describes the present co-occurrence search user options of the SCI and Google, and presents the algorithms that we developed to provide a strict constrained co-occurrence search user option. A heuristic approach to determine optimal word spacing (which we have implemented in our text mining studies) is also presented.

1a. Present SCI Co-Occurrence Search User Options

This document is based on the journal paper "RN Kostoff, JT Rigsby, and RB Barth. Adjacency and proximity searching in the Science Citation Index and Google. Journal of Information Science. 32. In Press.". Also, based on the report "RN Kostoff, JT Rigsby, and RB Barth. Adjacency and proximity searching in the Science Citation Index and Google. DTIC Technical Report Number ADA 442 888 (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA. 2005."

The SCI contains two Boolean operators for co-occurrence search, AND and SAME. SAME provides more proximity control than AND, and SAME is defined as follows in the online SCI Help:

> "SAME - All terms separated by the operator must appear in the same subfield. A subfield is defined as: 1) a sentence ending with a period, as in an abstract; 2) a phrase or text string ending with a semi-colon, as in the Keywords Plus field, or 3) a text string ending with a line break, as in the Address field."

Thus, the SAME operator does not allow the more precise word separation control implied by the stricter definition of constrained co-occurrence searching. In practice, for very short semantic units as defined above, for relatively specific search words, and for semantic units that focus on one concept, differences between the SAME operator and strict constrained co-occurrence search may not be substantial. However, for long semantic units, relatively general search terms, or semantic units covering different concepts, differences could be substantial.

1b. Present Google Co-Occurrence Search User Options

Google does not contain a strict constrained co-occurrence search user option, as defined above. Google does allow for an AND operator (its default when words are entered into the search window without quotation marks), whereby Web pages that contain all the words are returned. Thus, if the Google search window contains 'information technology' not enclosed by quotes, the search operation defaults to looking for Web pages that contain 'information' and 'technology'. This unconstrained co-occurrence search is particularly egregious for Google, since Web pages can be very long, and contain myriad diverse concepts on one page.

Google does allow a word wildcard search. According to the Google Help function (Cheat Sheet) Operator Example, 'red * blue' would find pages containing "the words **red** and **blue** separated by one or more words". This is not a constrained co-occurrence search capability in the above sense, but rather the inverse of such a capability. Insertion of more than one asterisk between words A and B is not defined in the Google Help function.

Our asterisk insertion experiments confirmed the single asterisk insertion definition stated above, and showed that insertion of two asterisks between words A and B in the title displayed retrieved records where words A and B were separated by exactly two words in the title. However, only a handful of records were displayed out of the many thousands of records ostensibly retrieved. Thus, if 100000 records were listed as retrieved, perhaps four or five were displayed, as opposed to the hundreds that would ordinarily be expected for display from a Google search. Therefore, it cannot be stated with any certainty that only records with two word spacing between words A and B in the title were retrieved, because of the small sample displayed. Insertion of three asterisks between words A and B retrieved records with three words between A and B in the title, and again only a handful were displayed. However, the fact that some word placeholder capability was exhibited by Google offered hope that a constrained co-occurrence search capability could be generated.

2. Approach and Results

2a. SCI

We have developed a simple algorithm that allows strict constrained co-occurrence searching as a user option in the SCI. It is based on the use of stopwords in the SCI search engine, defined in the online Help function of the SCI as follows:

> "Stopwords are frequently used words such as articles, prepositions, and pronouns that are ignored in a search. You may include a stopword in a search phrase, but it will function as a placeholder. For example, the search **mind over matter** will retrieve records that contain the phrases *mind and matter* and *mind in matter*, as well as *mind over matter*."

In other words, intercalating stopwords into a query phrase will retrieve those records that contain the word components of the phrase separated by the number of stopwords, in at least one location. Thus, the query term 'NUTRIENT of of of UPTAKE' (where 'of' is an SCI stopword) will return those records where 1) Nutrient precedes Uptake in at least one location, and their separation in that location is precisely three words. More generally, if A and B are two query terms, and the proximity operation desired is A w/n B (where w is abbreviation for 'within', and each retrieved record will have A and B separated by n or less words in at least one location in the record), then the general algorithm for A w/n B can be written (by integration over phrases with n or less adjacent stopwords separating A and B) as:

AB OR BA OR (ASB) OR (BSA) OR (ASSB) OR (BSSA) …… OR (A[(n-1)S]B) OR (B[(n-1)S]A) OR (A[nS]B) OR (B[nS]A) **(Condition 1a)**

where S is any stopword.

If n=4, for example, then the expression (A[4S]B) is written as (ASSSSB).

As an example of the form of the proximity search algoritm in practice, if A is the word NUTRIENT* and B is the word UPTAKE* (where * denotes the wildcard), then the formula for NUTRIENT* w/4 UPTAKE* can be written:

NUTRIENT* UPTAKE* OR (NUTRIENT* of UPTAKE*) OR (NUTRIENT* of of UPTAKE*) OR (NUTRIENT* of of of UPTAKE*) OR (NUTRIENT* of of of of UPTAKE*) OR UPTAKE* NUTRIENT* OR (UPTAKE* of NUTRIENT*) OR (UPTAKE* of of NUTRIENT*) OR (UPTAKE* of of of NUTRIENT*) OR (UPTAKE* of of of of NUTRIENT*)

If a more restricted proximity query is desired, for example those records where A and B are separated by n words, and records with A and B separated by less than n words are excluded, then the search algorithm can be written:

(A[nS]B) OR (B[nS]A) NOT (AB OR BA OR (ASB) OR (BSA) OR (ASSB) OR (BSSA) …… OR (A[(n-1)S]B) OR (B[(n-1)S]A)) (**Condition 2a)**

This document is based on the journal paper "RN Kostoff, JT Rigsby, and RB Barth. Adjacency and proximity searching in the Science Citation Index and Google. Journal of Information Science. 32. In Press.". Also, based on the report "RN Kostoff, JT Rigsby, and RB Barth. Adjacency and proximity searching in the Science Citation Index and Google. DTIC Technical Report Number ADA 442 888 (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA. 2005."

The advantage of adding these proximity search user options to the SCI depends not only on the length of the semantic unit, but on the characteristics of the words/ phrases in the query. For example, consider two extreme cases: A and B are very general components that form a general phrase when adjacent (e.g., information technology, computer science), and A and B are very specific components that form a specific phrase when adjacent (e.g., nutrient* uptake, isotope separation). The following table shows the number of relevant records retrieved as a function of the separation distance (distance between A and B in the query) for 'information technology' and 'nutrient* uptake'. To generate the matrix elements of this table, Condition 2a was used for the word pairs 'information technology' and 'nutrient* uptake*'. The separation distance between the words in each pair was varied from zero words (multi-word phrase) to eight words. For each value of separation distance, the query was run in the SCI search engine, and the ten most recent records retrieved were examined and judged for relevance to the context of the multi-word phrase (e.g., 'information technology'). The number of relevant records was entered into the appropriate matrix element. For example, the matrix element 'Information Technology-4', with an element entry of 3, means there were three relevant records (out of ten sampled) for the case where 'Information' and 'Technology' were separated by four words, and all records with 'Information' and 'Technology' separated by three or less words were excluded.

Table 1. Relevant Records vs Separation Distance

(ten records sampled for each matrix element)

Separation Distance (words) / **Adjacent Phrase**	0	1	2	4	8
Information Technology	10	5	9	3	2
Nutrient* Uptake*	10	10	10	10	8

The term with very general components ('information technology') yields substantial non-relevant records rapidly with increasing separation, whereas the term with very specific components ('nutrient* uptake') is more robust with respect to increasing separation distance. There was an anomaly in two word separation for 'information technology', because of the frequent appearance of the phrase 'Information and Communication Technology'. Similarly, almost all the records for 'nutrient uptake' separated by one word included the appearance of 'uptake of nutrients'. Appendix 1 contains an expanded version of this table.

Table 1 offers a heuristic approach to determining optimal spacing between search words for proximity or adjacency searching. Specifying the maximal spacing between words as the point where relevance fraction drops below some threshold will insure an appropriate balance between recall and precision. For the case of

'information technology', a maximum spacing ceiling of two is suggested by the table, whereas for 'nutrient uptake' a spacing ceiling of eight or greater may be appropriate.

Finally, it should be noted that the algorithm works on the Title field and the Topic search field (Title, Abstract, Keywords), but not the other fields (e.g., Author, Source, Address). Further, it works only in the same sentence, including across clauses, but may run into problems for sentences that contain internal periods. In our applications, neither of these limits has proven to be a problem.

2b. Google

We have developed a simple algorithm that allows strict constrained co-occurrence searching as a user option in Google. It is based on the use of the asterisk (*) in the Google search engine as defined above. A fixed spacing adjacency search algorithm is developed first, then variable spacing adjacency search algorithms and proximity algorithms are developed by integration over the fixed spacing adjacency search conditions, using spacing as the variable of integration.

If A and B are two words in the Google query, then the following fixed spacing adjacency conditions hold:

Condition 1b) Zero word spacing (coherent phrase) – "A B" –"A * B"
Condition 2b) One word spacing – "A * B" –"A * * B" (the minus sign is a NOT operator in Google)
Condition 3b) Two word spacing – "A * * B" –"A * * * B"
Condition 4b) Three word spacing – "A * * * B" –"A * * * * B"

and so on. For example, if the query "information * * technology" –"information * * * technology" is used to search the titles in Google, it will retrieve only those records that contain 'information' preceding, and separated by two words from, 'technology'.

In all these cases, when Google was searched, many thousands of records were retrieved, and many hundreds were displayed, as in a standard Google search. Additionally, in Condition 2b for example, the words A and B, separated by two asterisks, function to retrieve all records in which A and B are separated by two or more spaces, analogous to the Google-defined function of words A and B separated by one asterisk retrieving all records separated by one or more spaces. The same analogy holds true for word combinations separated by three spaces, four spaces, and so on.

Thus, to obtain the variable spacing adjacency search algorithm, integrate over the appropriate conditions above. For example, to obtain the adjacency search algorithm for A ADJ/2 B, sum over the first three conditions. To obtain the proximity search algorithm for A w/2 B, sum the first three conditions above and their analogs with B and A switched.

This document is based on the journal paper "RN Kostoff, JT Rigsby, and RB Barth. Adjacency and proximity searching in the Science Citation Index and Google. Journal of Information Science. 32. In Press.". Also, based on the report "RN Kostoff, JT Rigsby, and RB Barth. Adjacency and proximity searching in the Science Citation Index and Google. DTIC Technical Report Number ADA 442 888 (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA. 2005."

3. Conclusions

Simple algorithms that will allow strict constrained co-occurrence searching (adjacency and proximity) as user options in the SCI and Google have been developed. The SCI algorithm's value over use of the SAME operator, which requires co-occurrence of the search terms only in the same semantic unit (e.g., Abstract sentence), depends on the length of the semantic unit, the diversity of topics covered in the semantic unit, and the specificity of the search terms. The longer the semantic unit, the more diverse the topics covered, and the more general the search terms, the greater the value of the developed adjacency search algorithm. For applications such as literature-based or literature-assisted discovery [6], where queries developed initially for retrieving a target literature (e.g., water purification) are expanded and generalized to retrieve literatures directly and indirectly-related to the target literature (e.g., mass separation), the addition of adjacency or proximity searching offers major benefits compared with use restricted to the present SCI Boolean operators.

The Google algorithm's value over the default use of the AND operator, which requires co-occurrence only in the same Web page, depends on the three parameters mentioned above for the SCI, and is perceived to be even greater than that for the SCI because of the added length and diversity of Google Web pages compared to SCI journal paper Abstracts.

4. References

[1] RN Kostoff. Text mining for global technology watch. *Encyclopedia of Library and Information Science*. Second Edition. Drake, M., Ed. Marcel Dekker, Inc. New York, NY. Vol. 4. 2789-2799. 2003.

[2] RN Kostoff, RR Schaller. Science and technology roadmaps. *IEEE Transactions on Engineering Management*. 48(2). 132-143. 2001.

[3] RN Kostoff. Stimulating innovation. *International Handbook of Innovation.* Larisa V. Shavinina (ed.). Elsevier Social and Behavioral Sciences, Oxford, UK. 388-400. 2003.

[4]. MC McJunkin. Precision and recall in title keyword searches. *Information Technology and Libraries*. 14(3): 161-171. 1995.

[5] EM Keen. Some aspects of proximity searching in text retrieval-systems. *Journal of Information Science.* 18(2): 89-98 1992.

[6] RN Kostoff. Systematic acceleration of radical discovery and innovation in science and technology. Technological Forecasting and Social Change. In Press. Also, RN Kostoff. Systematic acceleration of radical discovery and innovation in science and technology *DTIC Technical Report Number ADA430720* (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA. 2005.

5. Disclaimer

The views in this Correspondence are solely those of the authors, and do not represent the views of the Department of the Navy, any of its components, or of DDL-OMNI Engineering, LLC.

6. Acknowledgements

We appreciate the review comments of Lawrence Barrett, Kathy Burrows, David Younghusband, and Teresa Loughbrough, of the Department of Knowledge and Information Science, Unilever Colworth Laboratory, UK.

7. Appendix 1 – Expanded Table of Relevance vs Separation Distance

Terms additional to those in Table 1 were examined for relevance and records retrieved as a function of spacing between words. The results are shown below, in Table 2.

Table 2 – Relevance and Number of Records vs Spacing between Words

(Ten Records Sampled for Each Spacing)

WORD COMBINATION/ SPACING (#WORDS)		0	1	2	3	4	5	6	7	8	9	10
COMPUTER SCIENCE												
	#REL/10	10	7	2	1	0	2	1	2	1	0	0
	#RECORDS	2763	33	135	62	53	54	26	28	28	27	28
ENERGETIC MATERIAL*												
	#REL/10	10	2	3	1	0	0	0	3	0	0	0
	#RECORDS	802	57	37	56	45	35	40	30	15	20	20
NECESSARY CONDITION												
	#REL/10	10	5	10	8	4	2	1	1	0	1	0
	#RECORDS	8937	1238	10377	1165	673	426	381	337	274	286	248
NAVIER STOKES												
	#REL/10	10	9	10	9	8	10	9	9	10	8	10
	#RECORDS	14786	78	15	15	9	14	9	19	13	8	15
MONTE CARLO												
	#REL/10	10	10	10	10	10	10	10	10	10	10	10
	#RECORDS	77516	19	59	53	43	56	56	49	47	53	42

This document is based on the journal paper "RN Kostoff, JT Rigsby, and RB Barth. Adjacency and proximity searching in the Science Citation Index and Google. Journal of Information Science. 32. In Press.". Also, based on the report "RN Kostoff, JT Rigsby, and RB Barth. Adjacency and proximity searching in the Science Citation Index and Google. DTIC Technical Report Number ADA 442 888 (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA. 2005."

In Table 2, ten records were sampled for each spacing, and the matrix entry in the row labelled #REL/10 signifies the number of relevant records, according to the judgment of the analyst. The matrix entry in the row labelled #RECORDS is the total number of records retrieved from the SCI using the two-word query.

The first three word combinations (COMPUTER SCIENCE, ENERGETIC MATERIAL*, NECESSARY CONDITION) are similar in word component generality to Information Technology (shown in the main text), and have a similar relevance-spacing profile. There is a rapid drop-off of relevance with spacing, as each component word finds a stronger association with other words in the semantic unit once the two words become separated. Additionally, in Table 1, there was a spike in records retrieved and relevance at a spacing of two between Information and Technology, due to the existence of the commonly used phrase Information and Communications Technology. There is a similar spike in relevance and records retrieved at a spacing of two between Necessary and Condition, due to the existence of the commonly used phrase Necessary and Sufficient Condition. Finally, there is a spike in records retrieved, but not relevance, at a spacing of two between Computer Science. In this case, there is no one dominant commonly used phrase with the spacing of two, but rather a number of lower frequency commonly used (but not necessarily related) phrases with a spacing of two, which in aggregate produce the spike in numbers of records retrieved.

The last two word combinations (NAVIER STOKES, MONTE CARLO) are similar in word component specificity to Nutrient* Uptake* (shown in the text), and have a similar relevance-spacing profile. As spacing increases, they do not form strong associations with other words, but rather retain their strong association with each other.

As discussed above, the number of records retrieved tends to be large when a given spacing results in a commonly used phrase(s). Thus, in Table 2, since the two word phrases with zero spacing were known commonly used phrases, they have relatively large values at zero spacing. As the spacing gets large, the number of records retrieved approaches a stable value (although not always monotonically; there are oscillations around the trend line in many cases above), with gradual decay (again, not monotonically in every case). The stable asymptote results from the approximate randomness of the appearance of the two words in the text, and the approximate decay results from the spacing being larger than some of the semantic units and subsequent loss of these records.